Robust statistical modeling improves sensitivity of high-throughput RNA structure probing experiments

Robust statistical modeling improves sensitivity of high-throughput RNA structure probing experiments

Alina Selega[1], Christel Sirocchi[2], Ira Iosub[2], Sander Granneman[2], Guido Sanguinetti[1,2]

[1] *University of Edinburgh, School of Informatics, EH8 9AB, Edinburgh, UK*

[2] *University of Edinburgh, Centre for Synthetic and Systems Biology, EH9 3BF, Edinburgh, UK.*

Structure probing coupled with high-throughput sequencing holds the potential to revolutionize our understanding of the role of RNA structure in regulation of gene expression. Despite major technological advances, intrinsic noise and high coverage requirements greatly limit the applicability of these techniques. Here we describe a probabilistic modeling pipeline which accounts for biological variability and biases in the data, yielding statistically interpretable scores for the probability of nucleotide modification transcriptome-wide. We demonstrate on two yeast data sets that our method has greatly increased sensitivity, enabling the identification of modified regions on many more transcripts compared with existing pipelines. It also provides confident predictions at much lower coverage levels than previously reported. Our results show that statistical modeling greatly extends the scope and potential of transcriptome-wide structure probing experiments.

RNA structure plays a key role in regulating RNA stability, transcription, and mRNA translation rates. In order to identify novel RNA structural regulatory elements, chemical and enzymatic structure probing is routinely used to interrogate RNA structure both *in vivo* and *in vitro*[1]. Current *in silico* RNA structure prediction programs rely on thermodynamic estimates to generate the most

likely secondary structure models. By incorporating data from structure probing experiments, the accuracy of secondary and tertiary RNA structure prediction can be significantly improved[2,3].

Most chemical RNA structure probing methods rely on the formation of adducts or cleavage of the RNA backbone, using as probes dimethylsulfate (DMS) and SHAPE reagents such as 1M7 (1-methyl-7-nitroisatoic anhydride) and NAI (2-methylnicotinic acid imidazolide)[4,5]. In all of these methods, the reagents terminate reverse transcription (RT), enabling detection of the sites of cleavage or modification by primer extension analyses, followed by mapping the RT drop-off position back to the reference sequence. These methods can be combined with next-generation sequencing (NGS) to simultaneously probe thousands of RNA molecules, as well as very long RNAs, in a single RT reaction. Insights obtained by these techniques include the largely unstructured state of stress-responsive transcripts in yeast and plants[6,7]. Recently, we developed the ChemModSeq structure probing pipeline to gain deeper understanding of RNA structural changes in long ribosomal RNA precursors during ribosome assembly[8].

NGS is certainly revolutionizing the RNA structure probing field, however, several data analytic questions need to be addressed. Firstly, NGS is often plagued by sequencing representation and coverage biases introduced during library preparation[9]. Identifying and correcting such biases is essential to avoid erroneous interpretations, however, to our knowledge, current methods do not address these issues. Secondly, statistical assessments must be informed by an analysis of inter-replicate variability in both control and treatment samples. Except for Mod-seq[10], current methods do not exploit replicate information, and, as a result, their output scores are not readily statistically

interpretable, often requiring setting arbitrary thresholds and other post-processing. Finally, a major question in the field concerns the coverage per nucleotide necessary to get reliable chemical reactivity values. Partly as a result of unresolved statistical issues in handling variability, current recommendations indicate that very high coverage levels are required[10,11], which are normally only met for a handful of transcripts in transcriptome-wide experiments.

To tackle these important issues, we developed *BUM-HMM* (Beta-Uniform Mixture Hidden Markov Model), a statistical machine learning pipeline for modeling NGS RNA structure probing data. BUM-HMM uses inter-replicate variability to identify transcript regions that are significantly more modified, incorporating coverage and sequence bias information within the model. The output of BUM-HMM is probabilistic, giving a transparent statistical interpretation which obviates the need for arbitrary thresholds and post-processing. We demonstrate that BUM-HMM is highly sensitive and remarkably robust even at low coverage, greatly improving over existing bioinformatic pipelines.

Results

To demonstrate the strength of the BUM-HMM method, we re-analyzed high-throughput DMS and 1M7 RNA structure probing experiments performed on yeast 40S ribosomes[8]. This study generated biological triplicates of each chemical probing experiment with very high sequence coverage, both in treatment and control samples (Supplementary Table 1). As secondary structure models for rRNAs and crystal structures of yeast ribosomes are now readily available[12,13], these data allowed

us to investigate the sensitivity and specificity of BUM-HMM compared to existing methods. In addition, we also generated two *in vivo* yeast mRNA transcriptome data sets using NAI as chemical probe (see Methods for details), which enabled us to test the performance of BUM-HMM in the important context of a transcriptome-wide mRNA structure probing experiment. For these analyses, between 36 and 55 million paired cDNA sequences were analyzed per sample (see Supplementary Table 1 and Methods for details).

Data preparation and model

All cDNA libraries were generated by random priming[6,8,11,14] and paired-end sequenced (see Methods and Supplementary Fig. 1 for details). Paired-end sequencing allows normalization for different read depths through calculating drop-off rates, which we define as the total number of reads stopping at a nucleotide divided by the total number of reads that cover that nucleotide[8,14]. The full procedure is described in detail in Methods and schematically illustrated in Fig. 1.

Briefly, we quantify biological variability using the log-ratio between the drop-off rates at the same nucleotide in a pair of control replicates (*log dor ratio*, LDR), for all possible pairs. We assemble all control LDRs in a null distribution (Step A) and correct sequence and coverage biases (Step B) to control for confounders (see Methods and Supplementary Fig. 2 for details). We then evaluate empirical p-values for all treatment-control LDRs at each nucleotide (Step C) and model these p-values using a Beta-Uniform mixture hidden Markov model (Step D) with hidden states corresponding to presence or absence of modification (see Methods and Supplementary Fig. 3 for a theoretical justification of the Beta-Uniform choice). We use BUM-HMM to compute

posterior probabilities of chemical modification for all nucleotides (Step E), providing a robust and statistically interpretable readout.

It is important to remark at this stage that, while single molecules are either modified or not at a particular locus, interpreting structure probing data as binary may appear overly simplistic. Transcripts *in vivo* exhibit dynamic secondary structures and may be bound by different proteins, so that different molecules of the same transcript may be accessible to chemical reagents at different positions. Furthermore, not all accessible nucleotides will be modified at low reagent concentrations, such as those typically used in structure probing experiments. The correct interpretation of the probabilistic output of BUM-HMM is therefore not that all transcript molecules with high posterior probability at a locus are in a specific state of accessibility, but that the proportion of modified molecules is sufficiently large to lead to an LDR value which cannot be explained by random variability alone.

Performance comparisons

Interpreting and evaluating the outcome of structure probing experiments is a notoriously difficult task due to a lack of "ground truth" examples to validate model predictions (see also Discussion). In this respect, yeast 18S ribosomal RNA represents an important case of a high abundance transcript with a well-defined and very stable secondary structure. We therefore first evaluated BUM-HMM's performance in terms of recovering the 18S structure from a recently published chemical probing data set[8]. These data sets have extremely high coverage (with a mean coverage per nucleotide close to 1 million for some samples, Supplementary Table 1), which clearly cannot

be achieved on many transcripts in transcriptome-wide studies. We thus later examine the performance of BUM-HMM on the transcriptome data set, which reflects a more realistic coverage scenario. We demonstrate through a number of case studies how BUM-HMM can aid the use of structure prediction algorithms and recover structural features in conserved areas of transcripts, as well as examine the robustness of BUM-HMM towards variations in coverage.

BUM-HMM demonstrates state-of-the-art performance recovering the structure of 18S with readily interpretable output

Guided by the available 80S and 40S structures[12,13], we determined which nucleotides were accessible and single-stranded and should, in theory, be therefore modified by 1M7 or DMS. Notice that this crystallographic structure is different from the phylogenetic (predicted) structure used in other studies[15]. As DMS preferentially reacts with A's and C's, we were able to examine the sensitivity and specificity of BUM-HMM. From many existing bioinformatic approaches[6–8,10,14,16], we chose the following methods to compare our model to: structure-seq[6], ΔTCR[14], which was the strongest performer in a recent review[16], and Mod-seq[10], which to our knowledge is the only method supporting multiple biological replicates. We evaluated all methods using the *receiver operating characteristic* (ROC), which plots the false positive rate against the true positive rate for different discrimination thresholds. A random predictor would have the *area under the* ROC *curve* (the AUC statistic) equal to 0.5 and a higher value of the AUC indicates better performance. When evaluated against the known crystal structure, BUM-HMM and ΔTCR were clearly the best

performers with AUC of 0.73 and 0.74, outperforming structure-seq and Mod-seq scoring at 0.68 and 0.64, respectively. The 1M7 data set demonstrated similar performance between methods (Supplementary Table 2).

However, the dynamic output ranges of the methods vary dramatically; to enable comparisons with BUM-HMM while taking into account these differences, we separately examined the true positive and true negative rate for different discrimination thresholds (scaling the scores to range between 0 and 1). BUM-HMM demonstrated a 20% increase of the true positive rate throughout most of the dynamic range compared to the other methods, for only a small decrease of the true negative rate (Fig. 2a and 2b).

Fig. 2c shows the proportions of nucleobases called as modified by all methods, when discriminating the scores at low, medium, and high thresholds or considering all scores greater than zero. BUM-HMM has excellent specificity to A's and C's throughout its dynamic range. On the contrary, structure-seq and ΔTCR do not discriminate as well between C's, G's, and U's when considering all scores, demonstrating their reliance on arbitrary thresholds as the means to remove noise. BUM-HMM identifies over a hundred modified nucleotides with very high posterior probabilities, many more than the other methods do when considering high reactivity thresholds. It is interesting to observe that on the 18S DMS data, BUM-HMM generates an almost binary output, with few values between 0 and 1. This reflects the stability of the 18S transcript clearly evident from the data, rather than a property of the model: BUM-HMM generates many more intermediate values on the transcriptome data set.

Fig. 2d shows a fragment of the 18S secondary structure as predicted by BUM-HMM, with many single-stranded A's and C's correctly identified. The results for all methods are shown on the 18S secondary structure models in Supplementary Fig. 4.

BUM-HMM output aids computational prediction of secondary structures

As explained earlier, the output posterior probabilities of BUM-HMM should not be directly interpreted as secondary structure readouts in general. These probabilities can, however, provide valuable constraints to energy-based structure prediction software, such as RNAstructure[17], ViennaRNA[18], and others. Such software predicts secondary structures of transcripts by minimizing the free energy associated with a particular "sequence–structure" configuration. For all but the shortest transcripts, this is a very difficult combinatorial optimization problem, resulting in many nearly equivalent optima corresponding to different structures. Transcripts *in vivo* are highly dynamic and can therefore exist in many different such configurations. However, under physiological constraints, it can be expected that only a subset of all possible structures (from a free energy point of view) will be present. We therefore used the BUM-HMM output as constraints for structure prediction with the RNAstructure Web Server[17].

To quantify the improvement provided by the BUM-HMM constraints, we selected as representative examples the SCM4, RPL37A, and RPL19B genes (coding sequence only), which encode a mitochondrial outer membrane protein and ribosomal 60S subunit proteins, correspondingly. These genes all have good coverage levels (mean coverage per nucleotide 799, 38711, and

15798), thus avoiding problems with missing information; they are also relatively long transcripts (564, 260, and 568 nucleotides long), and hence challenging for structure prediction algorithms. We used the *Fold*[17] method in RNAstructure to predict their secondary structure, with and without the BUM-HMM constraints. *Fold* returns an ensemble of generally around 20 low free-energy structures and we quantify the distance between two structures by using the binary Hamming distance. Constraining the algorithm with the BUM-HMM output considerably narrowed down the search space for free-energy minimization, as demonstrated by smaller Hamming distances between the resulting structures (Fig. 3a, 3b, and 3c). Further, these structures were more similar to the output of the alternative method *MaxExpect*[17] compared to only using sequence (Supplementary Fig. 5). We conclude that using posterior probabilities generated by BUM-HMM as algorithm constraints can improve secondary structure prediction for relatively long transcripts.

BUM-HMM correctly predicts structure of conserved regions in U3 snoRNA

While transcripts may co-exist in several different structural configurations, it is likely that some of their sections present increased structural stability for correct cellular functioning (e.g. in order to be bound by proteins). It is reasonable to expect highly conserved regions of a transcript to represent its more stable parts. To validate our model in the scenario of a more realistic transcriptome-wide coverage, we turned to the small nucleolar RNA U3. U3 is a model gene for evolutionary fitness studies[19] and has an accepted secondary structure in yeast[20], making it a good candidate for validation.

Even though the coverage on U3 was uneven and did not allow structural predictions on the whole molecule, BUM-HMM achieved the AUC of 0.76 when evaluated on the highly conserved regions located in boxes A, A', B, C, C', and D. Furthermore, when considering the longest conserved region with 16 nucleotides (box A and one highly conserved upstream nucleotide), BUM-HMM demonstrated excellent prediction accuracy of 0.88.

BUM-HMM has increased informativeness on transcriptome-wide analysis of RNA structure probing data

To evaluate the applicability of the methods in the transcriptome-wide scenario, we generated synthetic data sets by randomly selecting subsets of reads from the 18S DMS data set and evaluated the consistency of the methods at lower coverage (see Methods for full details). BUM-HMM showed excellent consistency as the mean coverage along the transcript was progressively reduced (Fig. 4a), retaining significantly above random accuracy even at a reduction of almost 2000 times (Supplementary Fig. 6). This performance challenges recent recommendations for the minimum coverage level for chemical probing experiments[11], indicating that BUM-HMM can obtain reliable predictions on a large fraction of transcripts in a standard transcriptomic experiment. Mod-seq and structure-seq exhibited considerably lower levels of consistency (Fig. 4c and 4d) and behaved as random predictors at the lowest coverage level. Highly consistent reactivity scores generated by ΔTCR (Fig. 4b) were largely due to its extreme conservatism at the chosen threshold of 50% of the dynamic range, at which it called no more than 20 nucleotides at all coverage levels. Notably,

all methods identified fewer modified nucleotides than BUM-HMM both on the full data set and at all coverage levels, this difference being particularly striking with ΔTCR and Mod-seq (Fig. 4b and 4c).

While performance analysis is hampered by a lack of a "ground truth" for most transcripts, a more general assessment of the informativeness of the methods' outputs is possible and instructive. We therefore quantified how many transcripts had at least 5% of their length called as modified by BUM-HMM and ΔTCR. We considered to be "called as modified" those nucleotides which obtained a score above 50% of the dynamic range of the model (having removed outliers for ΔTCR). With this procedure, BUM-HMM identified 2219 transcripts, while ΔTCR only retrieved 285. The low number of transcripts identified by ΔTCR is at odds with previous studies[6,7] suggesting that many RNAs are largely accessible and unstructured *in vivo*; this conservativeness may be due to the normalization procedures of ΔTCR[14] (see Supplementary Fig. 7 for illustration of associated problems).

We next analyzed the distribution of posterior probabilities across those mRNA transcripts which had a non-zero score attached to more than 75% of their length, which we call effectively probed. BUM-HMM selected 363 mRNA genes (Fig. 5a), which is in striking contrast with ΔTCR's 43 selected transcripts. When relaxing this criterion to (still highly informative) effective probing of more than 50% of the length, the number of mRNAs selected by BUM-HMM increased dramatically to 1764. Analyses of the 363 selected genes revealed that many appeared to have long segments of almost completely unstructured regions (such as TDH3, Fig. 5b) and many had

significant structure in the coding sequence (such as YOR365W, Fig. 5b). We next calculated the average FPKMs for these genes using the read counts from the control and treated sequencing data. This revealed a broad distribution with a median 191 (Fig. 5b) and the lowest FPKM of 60 (YOR385W, Fig. 5b and 5c). This gene had an average coverage of 335 reads per nucleotide, which we propose can be an indicative guideline of the lower bound on coverage required for high-throughput RNA structure probing experiments to effectively probe long transcripts.

Metabolic transcripts are generally flexible around the translation start site

Structure in untranslated regions (UTR) and around the translation start site (AUG) can reduce translation efficiency[21,22]. Recent high-throughput RNA structure probing also revealed a weak but significant negative correlation between RNA structure at that AUG *in vitro* and ribosome occupancy[23]. To test whether RNA structure measured *in vivo* also correlates with ribosome occupancy, we plotted the distribution of posterior probabilities around the translation start sites and performed a k-means clustering to identify patterns in the data. This revealed five clusters with different reactivity profiles (Fig. 5d). For the majority of transcripts, the region around the AUG had high posterior probabilities and therefore appeared to be largely unstructured (genes in clusters 0, 2, 3, 4). Interestingly, KEGG pathway analyses revealed that these clusters were highly enriched for transcripts encoding for ribosomal and metabolic proteins, in particular proteins involved in glycolysis/gluconeogenesis and amino acid biosynthesis (Supplementary Table 3). Remarkably, the more structured transcripts in cluster 1 were mostly enriched for transcripts encoding proteins

involved in mitochondrial translation (Supplementary Table 3).

One possible explanation for why the metabolic transcripts appear largely unstructured *in vivo* could be because they were occupied by ribosomes, which have an intrinsic RNA helicase activity to unfold structured regions within mRNAs[24]. We therefore asked whether there was a significant correlation between RNA flexibility within that region and ribosome occupancy on the transcripts. To test this, we calculated log_2 of the sum of posterior probabilities within 50 nucleotides around the AUG and compared it to the translational efficiency obtained from the recently published polysome microarray data[25] (Fig. 5e). This revealed that flexibility around the AUG did not positively correlate with polysome occupancy (Pearson correlation: -0.196, p-value = 0.0014). Similar results were obtained when using the entire 5' UTR region (Fig. 5f). Taken together, these results suggest that high ribosome occupancy alone is not sufficient to explain why certain transcripts were highly flexible in our *in vivo* NAI chemical probing data.

Discussion

High-throughput probing of RNA secondary structure offers unprecedented opportunities to elucidate the role of RNA structure in many fundamental biological processes. While the experimental platforms are rapidly reaching maturity, several data analytic issues hinder their wider applicability and adoption.

Our statistical pipeline addresses a number of such important problems. Firstly, it explicitly models the biological variability of the data, providing a statistical basis for determining the

significance of the observed signal. As such, it removes the need to set arbitrary thresholds and perform extensive post-processing of the analysis results, yielding a clean and statistically interpretable pipeline. This is in contrast to most existing methods and is a direct consequence of the probabilistic formulation of BUM-HMM. In this respect, it is indebted to earlier probabilistic models of SHAPE-Seq data[26]; notably, however, recent developments in the experimental technology, and in particular, the shift to random-primed experimental designs, force a major change in model architecture and motivate the non-parametric approach we take.

Our analysis identified important biases in the technology, especially prominent transcriptome-wide, which can have severe downstream consequences in any analysis. While random-priming designs effectively resolve the 3' biases of earlier SHAPE technologies, significant sequence and coverage biases remain. Our method provides automated empirical strategies for correcting these biases, potentially greatly extending the applicability of the technology.

Finally, the BUM-HMM model generates accurate and more informative results compared to other methods. Crucially, its predictions remain consistent with reduced coverage, demonstrating that the choice of an appropriate modeling framework can greatly increase the robustness of the technology. This is borne out by the effectiveness of BUM-HMM on a transcriptome data set with relatively low coverage: while current state-of-the-art methods can only provide information over a handful of transcripts, BUM-HMM selected more than 360 transcripts, some of which had a per nucleotide coverage as low as 335, heralding the advent of truly transcriptome-wide structure probing experiments.

While BUM-HMM addresses many of the data analytic challenges associated with structure probing data, it is important to stress that significant issues remain unsolved with the interpretation of such data. Many factors may affect accessibility (protein binding being a prime example), and in general transcripts *in vivo* may co-exist in multiple configurations, cautioning against simplistic interpretations in terms of secondary structure. How structure probing data may be used to inform model-based structure prediction is an important and active research field[27,28]. Our results show that BUM-HMM constraints, when incorporated in structure prediction algorithms, lead to more consistent structure models for many transcripts, demonstrating the importance of statistically sound data analytic strategies for downstream analyses.

Accession codes. The 18S rRNA (DMS and 1M7) and transcriptome-wide chemical probing sequencing data are available in the Gene Expression Omnibus under accession numbers GSE52878 and GSE78208, respectively.

Code availability. All of the code used in this study can be accessed in the following BitBucket repository: `https://bitbucket.org/aselega/bum_hmm_pipeline`. The BUM-HMM pipeline will be made available as a Bioconductor software package in due course.

Acknowledgements We thank all the members of the Granneman and Sanguinetti labs for critically reading the manuscript. This work was supported by grants from the Wellcome Trust to S.G. (091549) and I.I. (102334), a European Research Council grant to G.S. (MLC306999) and the Wellcome Trust Centre for Cell Biology core grant (092076). A.S. is supported by grants from the UK Engineering and Physical Sciences Research Council, Biological Sciences Research Council, and the UK Medical Research Council

(EP/F500385/1 and BB/F529254/1). Next generation sequencing was carried out by Edinburgh Genomics, The University of Edinburgh. Edinburgh Genomics is partly supported through core grants from NERC (R8/H10/56), MRC (MR/K001744/1) and BBSRC (BB/J004243/1).

Contributions All authors contributed to planning the experiments and computational procedures. C.S., I.I., and S.G. carried out the experiments. G.S. and A.S. developed the computational analysis pipeline. A.S., C.S., S.G., and G.S. performed the bioinformatics and computational analyses of the sequencing data. All authors contributed to writing the manuscript and approved the final manuscript.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to:
S.G. sgrannem@staffmail.ed.ac.uk.
G.S. gsanguin@inf.ed.ac.uk.

References

1. Kubota, M., Tran, C. & Spitale, R. C. Progress and challenges for chemical probing of RNA structure inside living cells. *Nature chemical biology* **11**, 933–941 (2015).

2. Wu, Y. *et al.* Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic acids research* **43**, 7247–7259 (2015).

3. Ouyang, Z., Snyder, M. P. & Chang, H. Y. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome research* **23**, 377–387 (2013).

4. Mortimer, S. A. & Weeks, K. M. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *Journal of the American Chemical Society* **129**, 4144–4145 (2007).

5. Spitale, R. C. *et al.* RNA SHAPE analysis in living cells. *Nature chemical biology* **9**, 18–20 (2013).

6. Ding, Y. *et al.* In vivo genome-wide profiling of rna secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).

7. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–705 (2013).

8. Hector, R. D. *et al.* Snapshots of pre-rRNA structural flexibility reveal eukaryotic 40S assembly dynamics at nucleotide resolution. *Nucleic acids research* gku815 (2014).

9. van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: tone down the bias. *Experimental cell research* **322**, 12–20 (2014).

10. Talkish, J., May, G., Lin, Y., Woolford, J. L. & McManus, C. J. Mod-seq: high-throughput sequencing for chemical probing of rna structure. *rna* **20**, 713–720 (2014).

11. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature methods* (2014).

12. Ben-Shem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science (New York, N.Y.)* **334**, 1524–1529 (2011).

13. Aylett, C. H. S., Boehringer, D., Erzberger, J. P., Schaefer, T. & Ban, N. Structure of a yeast 40S-eIF1-eIF1A-eIF3-eIF3j initiation complex. *Nature structural & molecular biology* **22**, 269–271 (2015).

14. Kielpinski, L. J. & Vinther, J. Massive parallel-sequencing-based hydroxyl radical probing of RNA accessibility. *Nucleic acids research* **42**, e70 (2014).

15. Tang, Y. *et al.* Structurefold: genome-wide rna secondary structure mapping and reconstruction in vivo. *Bioinformatics* btv213 (2015).

16. Kielpinski, L. J., Sidiropoulos, N. & Vinther, J. Reproducible Analysis of Sequencing-Based RNA Structure Probing Data with User-Friendly Tools. *Methods in enzymology* **558**, 153–180 (2015).

17. Reuter, J. S. & Mathews, D. H. Rnastructure: software for rna secondary structure prediction and analysis. *BMC bioinformatics* **11**, 1 (2010).

18. Lorenz, R. *et al.* Viennarna package 2.0. *Algorithms for Molecular Biology* **6**, 1 (2011).

19. Puchta, O. *et al.* Network of epistatic interactions within a yeast snorna. *Science* **352**, 840–844 (2016).

20. Méreau, A. *et al.* An in vivo and in vitro structure-function analysis of the saccharomyces cerevisiae u3a snornp: protein-rna contacts and base-pair interaction with the pre-ribosomal rna. *Journal of molecular biology* **273**, 552–571 (1997).

21. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science (New York, N.Y.)* **324**, 255–258 (2009).

22. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 3645–3650 (2010).

23. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103–107 (2010).

24. Takyar, S., Hickerson, R. P. & Noller, H. F. mRNA Helicase Activity of the Ribosome. *Cell* **120**, 49–58 (2005).

25. Arribere, J. A., Doudna, J. A. & Gilbert, W. V. Reconsidering movement of eukaryotic mRNAs between polysomes and P bodies. *Molecular cell* **44**, 745–758 (2011).

26. Aviran, S. *et al.* Modeling and automation of sequencing-based characterization of rna structure. *Proceedings of the National Academy of Sciences* **108**, 11069–11074 (2011).

27. Deng, F., Ledda, M., Vaziri, S. & Aviran, S. Data-directed rna secondary structure prediction using probabilistic modeling. *RNA* (2016).

28. Eddy, S. R. Computational analysis of conserved rna secondary structure in transcriptomes and genomes. *Biophysics* **43** (2014).

Figure 1: Overview of the BUM-HMM computational analysis pipeline. **(a)** Null distribution of LDRs is computed for all pairs of control replicate samples, quantifying variability in drop-off rate observed by chance. **(b)** Coverage-dependent bias is corrected by applying a variance stabilization transformation. For transcriptome-wide data sets, different null distributions are computed for different nucleobase patterns to address sequence-dependent bias. **(c)** Per-nucleotide empirical p-values are computed for all pairs of treatment and control replicate samples, by comparing the corresponding LDRs to the null distribution. **(d)** BUM-HMM is run on p-values as observations, leaving out any nucleotides with missing data. **(e)** The output is a posterior probability of modification, ranging from 0 to 1, for each nucleotide included in the analysis.

Figure 2: BUM-HMM identifies many modified nucleotides of 18S ribosomal RNA with high accuracy and specificity. (a, b) True positive rate and true negative rate of all methods for reconstructing secondary structure of 18S rRNA, evaluated against the known crystal structure. **(c)** Base composition of called nucleotides for all methods, when considering scores greater than: a value close to zero (10^{-6}), a low reactivity threshold (0.1), a medium reactivity threshold (0.4), and a high reactivity threshold (0.85). **(d)** A fragment of the 18S secondary structure with bases colored according to the BUM-HMM posterior probability at the corresponding nucleotide position.

Figure 3: Using BUM-HMM output results in more consistent secondary structure pre-

diction. (a) Distribution of Hamming distances between all pairs of secondary structures ($n = 20$) predicted for SCM4 by *Fold* when using only sequence (blue) and adding the BUM-HMM output as constraints (red), and a fragment of the lowest free energy structure. **(b, c)** Same as in **(a)**, for RPL37A **(b)** and RPL19B **(c)**.

Figure 4: BUM-HMM is highly consistent at low coverage and calls more nucleotides modified. **(a)** Consistency of posterior probabilities generated by BUM-HMM on data sets with progressively lower mean coverage (shown on the x-axis), synthesized from the DMS data set for 18S rRNA (see Methods for details). For each coverage level, base composition of nucleotides called as modified is shown in a corresponding barplot, averaged across 10 selections of subsets. The barplot in a shaded rectangle corresponds to the base composition of called nucleotides on the full data set. **(b, c, d)** Consistency of reactivity scores generated by ΔTCR (**b**), Mod-seq (**c**), and structure-seq (**d**) on the same synthetic data sets, with prior outlier removal.

Figure 5: Flexibility of 5' UTR and ribosome occupancy do not show a significant positive correlation *in vivo*. **(a)** Distribution of posterior probabilities over 363 protein-coding transcripts. The heatmap displays posterior probabilities for 363 mRNA transcripts that were selected from the transcriptome-wide data. The mRNAs were sorted by length (from short to long) and extended at each end by 300 nucleotides. The two black lines indicate the position of the start codon and stop codon, respectively. **(b)** Genome browser examples showing posterior probabilities of a highly expressed gene (TDH3; average FPKM = 3491) and a lowly expressed gene (YOR385W; average FPKM = 60). **(c)** Violin plot showing the distribution of average FPKMs, calculated using

the sequence reads from the control and NAI datasets. **(d)** Many transcripts are flexible around the translation start site. The plot shows the distribution of posterior probabilities 50 nucleotides around the translation start site (AUG). K-means clustering revealed five clusters with different distributions of probabilities. On the right side of the heat map, cumulative plots are shown for each cluster. The number of genes (n) in each cluster is also indicated. **(e)** High structural flexibility does not correlate with high ribosome occupancy. For each gene, we calculated log_2 of the sum of posterior probabilities from the heat map data shown in (**d**) and plotted it against the log_2 of the reported enrichment of the transcript in polysomes[25]. **(f)** Same as in (**e**) but with the entire 5' UTR.

Methods

ChemModSeq library preparation.

The 18S DMS and 1M7 data sets were previously described[8]. To generate the NAI transcriptome-wide data set, yeast cells (BY4741 strain) were grown to exponential phase and harvested by centrifugation. Cells were subsequently resuspended in 1 volume of phosphate buffer saline (PBS). NAI (dissolved in DMSO) was added to the suspension in a final concentration of 100 mM (5% DMSO final) and incubated for 10 minutes at room temperature. Cells were harvested by centrifugation, washed with ice-cold PBS and snap-frozen in liquid nitrogen. Total RNA was extracted as previously described[29]. The mRNAs were isolated using the PolyATtract mRNA isolation kit, according to manufacturer's procedures (Promega). Two biological replicates were generated for the transcriptome-wide analyses. The ChemModSeq libraries were generated as

previously described[8]. Briefly, cDNA was generated by random priming using a random hexamer oligo[8]. Subsequently, a DNA adapter was ligated to the 3' end of cDNAs using CircLigase. These adapters contained a random nucleotide at the 5' end to minimize the sequence representation biases introduced during the linker ligation reaction. Following PCR, libraries were resolved on 2% Metaphor gels and fragments between 200-700 were gel purified. Samples were sequenced on Illumina HiSeq2500 systems.

Sequence data processing and raw data analysis.

To process the fastq files the pyCRAC package was used[30]. To demultiplex the raw sequencing data we used pyBarcodeFilter.py, after which the remaining random nucleotide was removed from the 5' end of the forward reads. The data were subsequently collapsed using pyFastqDuplicateRemover.py that utilizes the random barcode information present in the 5' adapters to remove potential PCR duplicates. The resulting fasta file was mapped to the *Saccharomyces cerevisiae* genome (version R64, ENSEMBL) using novoalign 2.05 and only uniquely mapped reads were considered. PyReadCounters.py was subsequently used to generate read counts and FPKMs for all annotated features. The resulting GTF output files were converted to tab-delimited files containing three columns: chromosome, genomic position, and coverage or drop-off counts using pyGTF2sgr.py. These files were then fed to the BUM-HMM model to generate posterior probabilities.

Data characterization.

Using the final output files (see Sequence data processing and raw data analysis), the drop-off rate was computed for all nucleotide positions in each replicate as a measure of nucleotide's reactivity to the probing reagent in a given experiment. By definition, the drop-off rate ranges between 0 and

1. All drop-off rates were normalized to a common median across replicate samples.

$$\text{drop-off rate} = \frac{\text{drop-off count}}{\text{coverage}}$$

A measure of inter-replicate variability at each nucleotide position is defined as the log-ratio of drop-off rates (LDR) in a pair of replicate samples i and j:

$$\log\left(\frac{\text{drop-off rate}_i}{\text{drop-off rate}_j}\right) = \log\left(\text{drop-off rate}_i\right) - \log\left(\text{drop-off rate}_j\right)$$

If the drop-off rates are similar in both samples, the LDR will be close to 0, indicating little variability. In contrast, different drop-off rates would result in an LDR large in absolute value. LDRs in control conditions collectively describe the variability in drop-off rates that could be observed in the absence of the probing reagent. The set of these define the *null distribution* of LDRs.

LDRs are then computed for each combination of treatment-control replicates, quantifying the difference between the drop-off rate observed in a treatment experiment with respect to a control replicate. These are compared to the null distribution giving rise to empirical p-values. For efficiency, LDRs are compared to the precomputed quantiles of the null distribution. The p-value of an LDR represents the probability of it being insignificantly different from what could be observed by chance.

$$p\text{-value} = 1 - q, \text{where } q \text{ is the closest quantile}$$

Preprocessing.

In order to use the log transform, it is necessary to ensure that no nucleotides have zero drop-off rates. Therefore, only those nucleotides with non-zero drop-off counts for a corresponding pair

of replicate samples are used. The pipeline also features a user-defined parameter describing the minimum level of coverage that nucleotides should have to be included in the analysis.

Model.

Empirical p-values, computed for each nucleotide position and each treatment-control comparison (of which there are nm for n treatment and m control experimental replicates) are passed onto a hidden Markov model. The model has a hidden state h_t ($t = 1...T$ for T nucleotides) representing the true binary state of the t-th nucleotide (modified, 1 or unmodified, 0) and the observed variable v_t, corresponding to the empirical p-value at that position. P-values corresponding to different pairs of treatment-control replicates are assumed to be independent measurements. Notice that, since p-values are used as features and not for decision making, no issues of multiple hypothesis testing arise.

Transition probabilities are defined through empirically derived lengths of single- and double-stranded stretches of nucleotides. The model assumes expected uninterrupted stretches of 20 double-stranded or constrained nucleotides and 5 single-stranded or flexible nucleotides.

Emission probabilities come from a Beta-Uniform mixture (BUM) model. This design exploits the result that p-values are uniformly distributed under the null hypothesis[31]. P-values corresponding to accessible nucleotides are modeled with a Beta distribution, which favors small values, accommodating the fact that accessible nucleotides would have LDRs greater than most values in the null distribution. The p-value distribution computed for the transcriptome-wide data set strongly agrees with this model (Supplementary Fig. 3). The HMM is run separately on con-

tinuous stretches of nucleotides with a user-specified minimum coverage threshold and a non-zero drop-off rate in at least one treatment sample.

$$p(v_t|h_t = 0) \sim U(0, 1)$$

$$p(v_t|h_t = 1) \sim Beta(\alpha, \beta), \text{with } \alpha = 1, \beta = 10$$

Optimization of parameters.

We provide a strategy to optimize parameters of the Beta distribution with respect to the data. This strategy uses the expectation-maximization (EM) algorithm[32] and Newton's optimization method.

The iterative EM-algorithm starts with the initial values of $\alpha = 1$ and $\beta = 10$, with which the posterior probabilities are computed. It then computes new estimates for α and β using Newton's optimization method. Newton's method finds the maximum of the expected complete data log-likelihood, or more precisely, its relevant terms. The shape parameters α and β only appear in the emission term and within that, only in the component corresponding to the modified state of the latent variable h_t.

The expected complete data log-likelihood is given by the following expression (all expectations are with respect to corresponding distributions):

$$\langle \log p(\nu_{1:T}, h_{1:T}|\alpha, \beta)\rangle = \langle \log p(h_1)\rangle + \langle \sum_{t=1}^{T}\sum_{n=1}^{N} \log p(v_t^n|h_t)\rangle + \langle \sum_{t=1}^{T-1} \log p(h_{t+1}|h_t)\rangle,$$

for $t = 1...T$ nucleotides and $n = 1...N$ number of treatment-control comparisons. The relevant term corresponds to emission probabilities (second term in the previous expression):

$$\langle \sum_{t=1}^{T}\sum_{n=1}^{N} \log p(v_t^n|h_t)\rangle = \sum_{t=1}^{T}\sum_{n=1}^{N} \log p(v_t^n|h_t = 0)p(h_t = 0|v_{1:T}^n) +$$

$$+\sum_{t=1}^{T}\sum_{n=1}^{N}\log p(v_t^n|h_t=1)p(h_t=1|v_{1:T}^n)$$

Within that expression, the relevant term corresponds to the modified state of the hidden variable (second term in the previous expression):

$$\sum_{t=1}^{T}\sum_{n=1}^{N}\log p(v_t^n|h_t=1)p(h_t=1|v_{1:T}^n)=\sum_{t=1}^{T}\sum_{n=1}^{N}\gamma_t\log\frac{(v_t^n)^{\alpha-1}(1-v_t^n)^{\beta-1}}{B(\alpha,\beta)}=$$

$$=F,$$

where $\gamma_t = p(h_t = M|v_{1:T}^n)$ is the responsibility.

The first order derivatives of F are:

$$\frac{\delta F}{\delta\alpha}=\sum_{t=1}^{T}\sum_{n=1}^{N}\gamma_t\log v_t^n-\gamma_t(\psi_0(\alpha)-\psi_0(\alpha+\beta))$$

$$\frac{\delta F}{\delta\beta}=\sum_{t=1}^{T}\sum_{n=1}^{N}\gamma_t\log(1-v_t^n)-\gamma_t(\psi_0(\beta)-\psi_0(\alpha+\beta))$$

The second order derivatives of F are:

$$\frac{\delta^2 F}{\delta\alpha^2}=\sum_{t=1}^{T}\gamma_t N(\psi_1(\alpha+\beta)-\psi_1(\alpha))$$

$$\frac{\delta^2 F}{\delta\alpha\delta\beta}=\sum_{t=1}^{T}\gamma_t N\psi_1(\alpha+\beta)$$

$$\frac{\delta^2 F}{\delta\beta^2}=\sum_{t=1}^{T}\gamma_t N(\psi_1(\alpha+\beta)-\psi_1(\beta)),$$

where ψ is the polygamma function. Log transform is applied at the beginning of the algorithm to ensure that the estimated α and β are positive. Posterior probabilities are recomputed with the new estimates of α and β and the process is repeated a maximum number of 10 times or until the parameter values stop changing within the small predefined tolerance range.

Bias correction.

We used the transcriptome-wide data set to identify potential confounding factors which influence the LDRs in the absence of a reagent. The aim is to transform all LDRs accordingly and eliminate the revealed biases.

Coverage bias.

The coverage bias was identified by plotting the control LDRs as a function of the inter-replicate mean coverage at the corresponding nucleotide position (Supplementary Fig. 2a and 2b).

This bias is corrected by learning the functional dependency between these variables and transforming the data to reduce the variance of LDRs. We model drop-off count as a binomially distributed variable, which thus has the following standard deviation:

$$\sigma[\text{drop off count}] = \sqrt{np(\text{drop off})(1 - p(\text{drop off}))}, \text{ for a nucleotide covered } n \text{ times.}$$

Consequently, LDR has a standard deviation of:

$$\sigma[\text{LDR}] \propto \frac{\sigma[\text{drop off count}]}{n} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Therefore, the functional relationship between log-ratios and coverage can be modeled as a $k\frac{1}{\sqrt{n}}+b$, with some unknown parameters k and b, which are learned from the data using a non-linear least squares technique. Then, all LDRs are rescaled by this model with fitted parameters. For efficient runtime on transcriptome-wide data sets, the LDRs are split in bins of equal coverage ranges and the 95$^{\text{th}}$ quantile of LDRs and mean coverage are computed for each bin. These are used for

parameter fitting. Supplementary Fig. 2c and 2d show that the transformed LDRs have reduced dependency on coverage.

Sequence bias.

We compared the resulting LDR null distributions when separately considering nucleobase patterns of three (AAA, AAT, AAG, ...). For each of the 64 combinations of nucleobases, the transcriptome sequence was searched for all places of its occurrence. The LDRs of the middle nucleotide at these occurrences defined the null distribution specific to this nucleobase combination. Supplementary Fig. 2e and 2f demonstrate significant differences between these null distributions.

To correct for this sequence-dependent bias, we store the quantiles of each of the 64 different null distributions and compute empirical p-values by keeping track of which nucleobase triplet corresponds to the current nucleotide position and looking up values from the corresponding null distribution.

Due to the short length of the 18S ribosomal RNA molecule, the sequence-bias correcting step was omitted from the analysis when handling the corresponding data sets.

Handling of missing data and outliers.

The methods used in the evaluation[6, 10, 14] not only generate scores with drastically differing dynamic ranges, but also assume different interpretations of the same score values. For instance,

ΔTCR makes no distinction between the equal drop-off rates in control and treatment conditions and no coverage, assigning a score of 0 in both cases. Structure-seq marks missing data with a dummy value, whereas Mod-seq clamps the scenarios of no coverage and no significant modification to the same score of 0. Further, the outputs of these methods have clear outliers, with a handful of values being much larger than the 99^{th} quantile of the output distribution. Therefore, simply choosing the midpoint of the dynamic range for binarizing the resulting classifications would result in as few as a single true positive for some methods.

Thus, when performing evaluation, we set the missing data (for those methods that use it) and the outliers (computed as the values greater than the 99.5^{th} quantile of the output distribution) to 0. Considering other strategies, such as removing outliers or only evaluating on the non-missing data, resulted in grossly limited outputs generated by some methods for the simulated low coverage levels. Our choice, while circumventing these problems and enabling comparisons, follows the commonly utilized assumption that the reactivity of zero does not carry significant structural information.

Overall, these difficulties expose the problems associated with the discussed methods; namely, the absence of a unified output scale (which therefore leads to arbitrary threshold setting), gross outliers, and inability to represent missing data, which thus results in extreme conservatism of the classification. BUM-HMM addresses this by having a clearly defined probability output range and separating out the nucleotides about which no predictions can be made.

When computing true positive and negative rates, the output scores of all methods were

normalized to the range of BUM-HMM. AUCs and true positive and negative rates were computed with the ROCR package[33]. When characterizing the methods' sensitivities using the DMS data set specific to A's and C's, the outputs of ΔTCR and Mod-seq were normalized with the 2-8% normalization rule[34] to enable comparisons at the same (previously used) low, medium, and high reactivity thresholds[34,35].

Secondary structure prediction.

When generating secondary structures informed by BUM-HMM, posterior probabilities were uploaded to the RNAstructure Web Server[17] as a SHAPE constraints file with default parameter values used. For RPL37A and RPL19B, the structure was predicted for the longest CDS region.

Performance evaluation of BUM-HMM on the conserved regions of U3 snoRNA.

Conservation scores associated with the human U3 snoRNA were taken from Rfam[36]. Highly conserved parts of the box regions, matching in sequence between the human[37] and yeast transcripts[20], were selected, with three lowly conserved nucleotides allowed in the middle of the regions (a total of 40 nucleotides). Evaluation was performed on those nucleotides with an attached posterior probability $p > 0$ (28 of those nucleotides).

582 **Lower coverage simulation analysis.**

583 To evaluate the output consistency of the methods at lower coverage levels, we generated synthetic
584 data sets by randomly selecting subsets of 2 million, 1 million, 100000, 30000, 20000, 10000, and
585 1000 reads from the 18S DMS data set. For each subset, 10 such selections were made. Files
586 with coverage and drop-off counts were generated for each selection and passed to BUM-HMM.
587 Consistency was evaluated with the AUC statistic between the output scores generated by each
588 method for a given synthetic subset selection and the whole data set. For all methods, outliers
589 were handled as described above and calling of modified nucleotides (used for the barplots of base
590 composition) was performed at the threshold of 50% of the dynamic range of each method, after
591 having dealt with the outliers.

592 **Methods References**

593 29. Tollervey, D. A yeast small nuclear RNA is required for normal processing of pre-ribosomal
594 RNA. *The EMBO journal* **6**, 4169–4175 (1987).

595 30. Webb, S., Hector, R. D., Kudla, G. & Granneman, S. PAR-CLIP data indicate that Nrd1-
596 Nab3-dependent transcription termination regulates expression of hundreds of protein coding
597 genes in yeast. *Genome biology* **15**, R8 (2014).

598 31. Murdoch, D. J., Tsai, Y.-L. & Adcock, J. P-values are random variables. *The American*
599 *Statistician* (2012).

600 32. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data
601 via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* 1–38
602 (1977).

603 33. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. Rocr: visualizing classifier performance
604 in r. *Bioinformatics* **21**, 7881 (2005). URL `http://rocr.bioinf.mpi-sb.mpg.de`.

605 34. Low, J. T. & Weeks, K. M. Shape-directed rna secondary structure prediction. *Methods* **52**,
606 150–158 (2010).

607 35. Lucks, J. B. *et al.* Multiplexed RNA structure characterization with selective 2'-hydroxyl acy-
608 lation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National*
609 *Academy of Sciences of the United States of America* **108**, 11063–11068 (2011).

610 36. Nawrocki, E. P. *et al.* Rfam 12.0: updates to the rna families database. *Nucleic acids research*
611 gku1063 (2014).

612 37. Granneman, S. *et al.* Role of pre-rrna base pairing and 80s complex formation in subnucleolar
613 localization of the u3 snornp. *Molecular and cellular biology* **24**, 8600–8610 (2004).

614 **Supplementary Information**

615 **Supplementary Table 1**: Overview of paired cDNA reads analyzed from each data set. All raw
616 sequencing data have been collapsed before aligning to the reference sequences to remove potential
617 PCR duplicates. Only properly paired reads were considered for the analyses.

Supplementary Table 2: Accuracy of reconstructing secondary structure of 18S ribosomal RNA from the 1M7 data set for all methods, measured with the AUC statistic against the known crystal structure of the rRNA.

Supplementary Table 3: KEGG pathway analysis of the k-means clusters shown in Fig. 4d. These analyses were performed on the string-db server (`www.string-db.org`).

Supplementary Figure 1: ChemModSeq library preparation design. Chemically probed RNAs were reverse transcribed with an oligonucleotide containing a random hexamer and an Illumina compatible sequence for PCR amplification. Subsequently adapters were ligated to the 3' end of cDNAs that contained six random nucleotides and a six nucleotide barcode followed by another random nucleotide. The latter was introduced to minimize sequence bias representation introduced during the CircLigase ligation reaction. The six random nucleotides were used to eliminate potential PCR duplicates. Indexing barcodes were added to the 3' adapter sequence by PCR. The in-read barcodes in the 5' end of the PCR product were processed using pyBarcodeFilter.py and reads were collapsed using pyFastqDuplicateRemover.py from the pyCRAC package[30].

Supplementary Figure 2: Coverage- and sequence-dependent biases were identified in the transcriptome data set. **(a, b)** Presence of a coverage-dependent bias, reflected by the dependency between the average LDR and the mean coverage at each nucleotide position in a pair of control replicate samples, for all such pairs. **(c, d)** Same dependency plotted as in (**a, b**) after applying a bias-correcting strategy to the LDRs. **(e, f)** Presence of a sequence-dependent bias, reflected by differing null distributions of LDRs, each computed only for nucleotide positions

corresponding to a given trinucleotide pattern.

Supplementary Figure 3: Distribution of empirical p-values for the transcriptome data set closely follows a Beta-Uniform distribution on both strands. The histograms show the distributions of empirical p-values associated with LDRs between all combinations of treatment and control samples on the transcriptome data set.

Supplementary Figure 4: BUM-HMM correctly identifies many flexible A's and C's as modified nucleotides. Secondary structures of 18S ribosomal RNA with bases colored according to the reactivity score or posterior probability at the corresponding nucleotide position, generated by BUM-HMM, ΔTCR[14], Mod-seq[10], and structure-seq[6] analysis pipelines.

Supplementary Figure 5: Using BUM-HMM output results in more consistent secondary structure prediction across different methods. (a) Distribution of Hamming distances between the structures predicted for SCM4 by *Fold*[17] ($n = 20$) and by *MaxExpect*[17] ($n = 3$ with sequence, $n = 1$ with BUM-HMM) when using only sequence (blue) and adding the BUM-HMM output as constraints (red). **(b, c)** Same as in **(a)**, for RPL37A **(b)** and RPL19B **(c)** (with *Fold*, $n = 20$ structures were generated, with *MaxExpect*, $n = 1$ structure).

Supplementary Figure 6: BUM-HMM retains good accuracy at 18S secondary structure reconstruction at lower coverage levels. Agreement with the 18S crystal structure of the posterior probabilities generated by BUM-HMM on data sets with progressively lower mean coverage (shown on the x-axis), synthesized from the DMS data set for 18S ribosomal RNA. For each

coverage level, the subsets of reads were randomly selected from the full data set 10 times.

Supplementary Figure 7: The ΔTCR algorithm produces very high numbers in regions with low coverage. Shown is a genome browser image of a gene (YHB1) with an FPKM of 190. The red-dotted box shows a region near the 3' end of the gene where there is low coverage. The top two panels show the ΔTCR[14] output, with the second panel displaying the same data but scaled to a maximum ΔTCR value of 0.025. The third panel shows the BUM-HMM posterior probabilities for the same region. The last four panels show the cDNA coverage over the gene from the two control RNA sequencing data and the two NAI treated sequencing data.